PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 4 :
C07K 15/00, C07H 15/12, 17/00
A61K 37/02

A1

(11) International Publication Number: WO 88/ 06601

(43) International Publication Date: 7 September 1988 (07.09.88)

(21) International Application Number: PCT/US88/00718

(22) International Filing Date: 2 March 1988 (02.03.88)

(31) Priority Application Number: 021,047

(32) Priority Date: 2 March 1987 (02.03.87)

(33) Priority Country: US

(71) Applicant: GENEX CORPORATION [US/US]; 16020 Industrial Drive, Gaithersburg, MD 20877 (US).

(72) Inventors: LADNER, Robert, Charles ; 3827 Green Valley Road, Ijamsville, MD 21754 (US). BIRD, Robert, E. ; 3903 Morrell Court, Kensington, MD 20895 (US).

(74) Agents: FOX, Samuel, L. et al.; Saidman, Sterne, Kessler & Goldstein, 1225 Connecticut Avenue, N.W., Suite 300, Washington, DC 20036 (US).

(81) Designated States: AT (European patent), BE (European patent), CH (European patent), DE (European patent), FR (European patent), GB (European patent), IT (European patent), JP, LU (European patent), NL (European patent), SE (European patent).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: GENE REPRESSORS

(57) Abstract

A gene repressor comprising two or more sequence-specific DNA-binding domains covalently linked by polypeptide and recombinant DNA molecule encoding same.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	ML	Mali
AU	Australia	GA	Gabon	MR	Mauritania
BB	Barbados	GB	United Kingdom	MW	Malawi
BE	Belgium	HU	Hungary	NL	Netherlands
BG	Bulgaria	IT	Italy	NO	Norway
BJ	Benin	JP	Japan	RO	Romania
BR	Brazil	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SN	Senegal
CH	Switzerland	LI	Liechtenstein	SU	Soviet Union
CM	Cameroon	LK	Sri Lanka	TD	Chad
DE	Germany, Federal Republic of	LU	Luxembourg	TG	Togo
DK	Denmark	MC	Monaco	US	United States of America
FI	Finland	MG	Madagascar		

TITLE OF THE INVENTION

GENE REPRESSORS

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to the field of gene repressors.

Brief Description of the Background Art

Genomic DNA is a two-stranded helix composed of four types of deoxynucleotides, deoxy-Adenine; deoxy-Thymine, deoxy-Cytosine, and deoxy-Guanine. These are abbreviated dA, dT, dC, and dG when RNA is also being discussed. If there is no chance of confusion, A, T, C, and G are sufficient.

RNA is frequently single-stranded and is composed of four types of nucleotides, Adenine, Uracyl, Cytosine, and Guanine. These are abbreviated A, U, C, and G. RNA serves several functions: messenger RNA (mRNA) is the template for protein synthesis, transfer RNA (tRNA) decodes the genetic code, ribosomal RNA (rRNA) is a structural component of ribosomes.

Proteins are composed of linear chains of amino acids. There are twenty types of amino acids encoded by genes. Proteins which catalyze chemical reactions are called enzymes. Proteins also perform mechanical and chemical functions.

Proteins carry out most catalytic functions in the cell as well as serving mechanical and chemical functions. DNA acts as the archive of genetic information. In cell division, the DNA is copied and each cell obtains a complete copy of the

entire genome. RNA serves as a working template for protein synthesis and other function.

Viruses are small packets of genetic material wrapped in protein and, sometimes, lipids and polysaccharides. The coat protein of a virus recognizes and binds to chemical features on the surface of suitable host cells. By one of a variety of mechanisms, the virus injects its genetic material into the host cell. In some viruses the genetic material is DNA, but in others it is RNA. RNA viruses contain a code for a special enzyme which will reverse the normal direction of transcription and make a DNA copy of the viral RNA. These are the so-called retroviruses.

Once the viral genome is inside the host as DNA, the virus may subvert the normal function of the cell and appropriate all of the cell's materials to synthesize more virus. This behavior leads to the death and rupture of the cell.

Alternatively, the viral DNA may become integrated into the host genome and cause no change in cell activity for a long time. Such silent viruses are termed temperate. Some future event may cause such a virus to reemerge and lead to rapid growth of virus progeny and cell lysis.

Viruses carry:

1) genes for their own structural proteins,
2) genes for proteins to control their own synthesis,
3) genes for proteins to control host activity, and
4) other non-essential genes.

The lambda phage is one of the best known temperate phages. Lambda phage grows in E. coli. Two classes of mutants of lambda are of special interest here. The first class is cI (c for clear) in which the gene for lambda repressor is absent or non-functional. The second class is vir (for virulent) in which the DNA to which lambda repressor binds is absent or non-functional. In both classes, lambda

immediately produces progeny and lyses E. coli. In class cI, supplying and expressing the repressor gene will prevent cell lysis. In class vir, however, no natural repressor will prevent cell lysis.

Among the many genes contained in lambda, two are particularly important here: CRO and repressor. If CRO is expressed, synthesis of repressor is blocked and lambda will grow in an actively lytic mode. If, on the other hand, repressor is expressed, synthesis of CRO is blocked and lambda will integrate into the E. coli genome and grow passively.

The bacteriophages P22 and 434 are closely related to lambda. They too possess CRO and repressor proteins; the exact sequences of these proteins and the DNA sequences to which they bind are different from lambda and from each other.

The chemical processes of a cell are regulated at several levels as illustrated in Fig. 2:

1) modulation of existing enzymes,
2) rate at which proteins are destroyed,
3) rate at which proteins are synthesized,
4) rate at which mRNA is destroyed, and
5) rate at which mRNA is created.

At the left of Fig. 2, DNA acts as the archive of genetic information. Repressors and positive regulators control the rate at which individual genes are transcribed into mRNA. The behavior of the repressors and positive regulators may be altered by interactions with effector molecules. In the center, mRNA may interact with effector proteins. The messages are translated by ribosomes into proteins. At a rate determined partially by its sequence, mRNA is degraded. On the right, proteins convert substrates into products. The behavior of proteins may be altered by interactions with effector molecules. Proteins, like mRNA have finite lifetimes.

The simplest regulation is to modulate the activity of existing proteins. Enzymes may be inhibited by their own products or by the product of a chain of reactions in which they play a part. If the product of an enzyme accumulates, it comes back to the enzyme and says, "Slow down, you're working too hard." Enzymes may be activated by chemicals which they act upon or by chemicals which will react with their product.

Proteins are degraded by proteases. The rate of degradation is determined by the amount of proteases in the cell and the susceptibility of each protein to each protease.

The concentration of enzymes may be regulated by the rates at which mRNAs are translated into protein. In some cases, proteins bind to the mRNA which codes for them, thereby reducing the rate of protein synthesis once enough protein is made. The genetic code is redundant; there are several codons for most of the amino acids (only met and trp have unique codons). Not all of the different codons for a given amino acid work equally well. Thus, a given gene will be expressed at a lower level if inefficient codons are used.

Different mRNAs have different life-times in the cell.

The ultimate level of control is that of transcription. If the gene is never transcribed into mRNA, the protein will never be made. It is this level of control to which the present invention is addressed.

The genome of an organism is divided into genes. Each gene contains a stretch of DNA which encodes for protein or for some classes of RNA such as tRNA or rRNA. In addition, there are control sequences which specify when the gene should be transcribed and how many copies of the RNA should be made.

These control sequences can be divided into two classes, positive control sequences and negative control sequences.

Positive control sequences bind proteins which can recognize specific DNA sequences. These proteins (called positive regulators) then allow RNA-polymerase (DNA-to-RNA

transcriptase) to bind and copy the following DNA into RNA. There may be only one or a few kinds of RNA-polymerase in a cell, but many classes of positive control sequences. The positive regulators act as decoders between DNA control sites and RNA polymerases. There are DNA sequences which bind RNA-polymerase without the assistance of positive regulatory proteins. These sequences are called promoters. Often positive regulatory proteins act to complement partial promoters.

Negative control sequences bind proteins which recognize specific DNA sequences. These proteins (called repressors) block the binding or passage of RNA-polymerase and thus prevent transcription of the gene. The negative and positive control sequences need not be disjoint. Indeed, positive regulators and repressors may compete for the same sites. In addition, one should remember that DNA is an antiparallel double helix and that valid genetic information may be read from either strand. Thus, a protein may be a repressor for a gene reading in one direction and a positive regulator for a different gene read in the other direction.

Bacterial and viral repressors are dimeric proteins. These molecules have been known to exist for over twenty years and an extensive literature describes:

a) the amino acid sequences of twenty or more of them,
b) the DNA sequences to which each repressor binds,
c) the relative affinity of different repressors for various DNA sites,
d) the domain structure of several repressors,
e) the three-dimensional structure of at least one repressor bound to DNA which it recognizes, and
f) the biological effect of the repressor.

The portion of the DNA to which the repressor binds is called an operator. Naturally occurring operators are approximately palindromic. Many repressors comprise not only

a dimeric DNA-binding domain but also a second dimeric domain which does not contact DNA. The second domain may recognize a chemical or physical agent and under certain conditions may cause the DNA-binding domain to change its ability to bind DNA. In these cases, repression is conditional. Some repressors bind DNA only in the presence of their signal chemical, others bind DNA only in the absence of their signal. In addition, the link between the DNA-binding domain and other parts of the repressor can be a site of action for a specific protease. Cleavage of the linker may irreversibly eliminate the ability of the recognition domain to dimerize or bind DNA.

Three-dimensional structures are known for the DNA-binding domains of three viral or bacterial repressors or positive regulators: lambda CRO, the amino-terminal domain of lambda repressor, and catabolite activator protein (CAP) of E. coli. Fig. 3 shows schematic representations of these three molecules.

Lambda CRO is the smallest, 66 amino acids. CRO forms dimers in solution. CRO binds to six similar, nearly palindromic sites in the genome of bacteriophage lambda. The CRO dimer contains two symmetry-related alpha helices which are properly positioned to fit into successive major grooves of B-DNA. Binding of CRO dimers to these six sites proceeds independently, i.e. the affinity of a CRO dimer for one site does not depend on whether any other sites are occupied. An X-ray structure of CRO and a model of how it might bind to DNA reveals that the amino-terminus of CRO is far from the DNA and that the carboxy-terminal probably does not wrap around the DNA. Mutants lacking a few residues at the C-terminus are functional.

Lambda repressor comprises 236 amino acids. The first 92 amino acids of each polypeptide chain in a dimer fold into a compact domain each of which contains an alpha helix. These helices are positioned very much like the corresponding

helices of CRO. Unlike CRO, the amino-terminal arms of the repressor probably do wrap around DNA adding both specific and non-specific binding. The carboxy-terminal parts of the chains contain two 38-amino-acid segments which seem to have no fixed conformation, followed by a pair of 104-amino-acid domains which dimerize strongly.

The 38-amino-acid linkers can be cut by mild proteolysis. The separate N-terminal domains do not dimerize at physiological concentrations, while the C-terminal domains do dimerize. This dimerization and weaker tetramerization seem to be the only functions of these domains.

The structure of the entire CAP molecule has been determined for the wild-type molecule and for a mutant which does not bind cAMP and always binds DNA. Normal CAP binds DNA only when cAMP is present. Although the positioning of two alpha helices in CAP are quite similar to those in CRO and lambda repressor, the direction of the helices is different. In CRO and lambda repressor, the C-terminal ends of the helices are closer together than are the amino-terminii, while in CAP the C-terminii are further apart then are the N-terminii.

The two-fold axes of the DNA-binding domain align with a two-fold axis of a palindromic sequence of DNA and equivalent amino acids of the repressor bind to equivalent portions of the DNA. Thus the requirement of a palindromic sequence is a direct consequence of the dimeric nature of the repressor. Even though many operators are only approximately palindromic, the dimeric nature of the repressors will not allow the operator to depart far from the palindromic paradigm. Dimeric repressors bind best to palindromic DNA.

The evolutionary reason for this seems clear. The backbone of DNA has a diad axis normal to the helix axis at every PO4 group and between each pair of PO4 groups. Dimeric proteins are common because each favorable mutation yields two

favorable interactions in the protein. A dimeric protein can easily provide interactions with the same symmetry as DNA. Dimeric proteins require less DNA, which is an important consideration for a virus. Viruses must be able to pack all their needed DNA inside their capsid. Indeed, viruses produce capsids with very high symmetry. The lower symmetry of DNA forces the virus and other organisms to use diadic proteins to interact with DNA.

For an organism, use of palindromic sequences in control regions is no serious restriction. Viruses often contain, however, control sequences very similar to host control sequences. What was needed was a means to disrupt the function of any virus without affecting host function. The present invention relates to designed repressors which bind to the unique asymmetric DNA which codes for viral proteins or RNAs, rather than to viral control sequences. In some viruses, the control sequences will be different from host control sequences. In these cases, symmetric repressors might be used. Even here, asymmetric repressors might be quite useful. As in lambda, viral controls may not be perfectly palindromic, a feature which the virus exploits to fine tune the strength of binding by its own positive regulatory proteins. To make repressors which will bind very strongly and prevent the virus' own proteins from binding the present invention provides adaptation to the existing asymmetry in the virus control sequences.

SUMMARY OF THE INVENTION

The present invention consists of methods to produce novel repressors which will bind to general (non-palindromic) sequences of DNA, the repressors generated by these methods, the genetic material which encodes these repressors, and the uses to which such repressors might be put.

The process can be divided into two major parts:

I. Devise a framework for sequence-specific DNA-binding proteins and create a dictionary of recognition elements;

II. Establish the specific recognition components for a selected DNA sequence.

Step I need be done only once for each repressor framework.

Step II needs to be done for each application.

The major applications of this invention are:

I. Therapy of Viral Diseases

II. Prevention of Viral Diseases

III. Control of Microorganisms

Viruses dump their DNA (or RNA) into the host cell. Viruses can be isolated and their DNA sequenced. Given the sequence of a virus, it is possible to find the open reading frames and the boundaries of the genes. Thus, one can select sites to repress and design and test repressors.

When a patient already has a viral disease, one must deliver the genes for repressors to as many infected cells as possible. This need for gene therapy goes beyond the scope of this application, but such methods are under development.

Certain viral diseases are well-known and affect animals and plants. Having developed repressors against these viruses, one can introduce the genes into the germ line and breed viral resistance into animals, plants, microbes, and even humans. These genes could be put under the same control as the same viral genes, so that they only turn on when there is a viral infection.

The use of switchable repressors will make control of microbes much easier.

Description of the Drawings

1 Transcription and Translation
2 Cellular Regulation
3 Natural DNA-binding Proteins
4 Natural Repressors and Operators
5 Cooperative Binding of Lambda Repressor
6 Development of General Gene Repressors
7 Schematic Asymmetric Bidentate Repressors
8 Schematic Tridentate Repressor Frameworks
9 Schematic Tetradentate Repressor Frameworks
10 Schematic Pentadentate Repressors
11 Natural & Created Bidentate DNA Binders
12 In vivo Selection for Heterodimers
13 Construction of Recognition Dictionary
14 In vivo Selection for Heterotetramers

DESCRIPTION OF PREFERRED EMBODIMENTS

The three-dimensional structures of repressors studied to date all contain a pair of symmetrically related alpha helices, as shown in Fig. 4, nine amino acids long. Fig. 4 shows schematically the faces for DNA and repressor which would bind together. Note that the DNA (above) is palindromic and that the protein has diad symmetry. The helices are positioned so that they will fit into the major groove of successive turns of B-DNA. The side chains of five of the amino acids in each of these helices make numerous contacts with edges of five base pairs of the DNA. There are some contacts with the bases on either side of the main five, and some amino acids outside the primary helices may contact the DNA. In the sequel, normal recognition will be described as contact between five base pairs and a protein helix of nine amino acids; parenthetical remarks will explain minor modifi-

cations needed when adjacent bases or other amino acids show noticeable interaction. The contact of a protein alpha helix with five base pairs will be taken as the unit of recognition: one such contact will be termed unidentate; two such contacts will be termed bidentate; three, tridentate, etc.

DNA recognition by protein is achieved by placing appropriate hydrogen-bonding and hydrophobic groups of the protein in correct relation to hydrogen-bonding and hydrophobic groups of the DNA. Non-specific interactions occur between protein and DNA backbone.

Eukaryotes have very much more DNA than do prokaryotes, about 1,000,000,000 base pairs in the human genome. The DNA in eukaryotic cells is mostly complexed with histones which act as universal repressors. Genes are expressed only when specifically activated by positive regulators. Transcription of activated genes could be repressed by a repressor which binds at or down-stream from the promoter site.

Structural and biochemical information strongly suggests that each recognition helix of a repressor recognizes five base pairs. Because of the dimerism, natural repressors contact ten base pairs with their recognition helices. To obtain correct geometry, these bases are divided into two sets of five separated by several (5-7) intervening bases. The identity of these intervening bases has little or no effect on the helical recognition process. These back-side base pairs may effect recognition by other components of the repressor (e.g., the amino-terminal arms of lambda repressor).

A typical bacterium has about 5,000,000 base pairs in its genome. If the genome is essentially random, we might expect each of the 1,049,576 different decanucleotides to appear about five times in the genome. Only 1024 of the different decanucleotides are palindromic. Actual repressors recognize more than ten bases by means of interactions which are not so easily described as the alpha helix-major groove combination.

For example, lambda repressor wraps arms around DNA and recognizes base pairs on the back side, while lambda CRO recognizes bases on the front side. The arms of lambda repressor are not resolved in the crystal structure of the isolated 92-amino-acid DNA-binding domain of repressor. No high-resolution X-ray structure of the complex between lambda repressor and its operator is available. The recognition obtained by the arm suggested from biochemical studies is between a lysine residue and a guanine on the back side. Because of the flexibility of the arm, this recognition is unlikely to distinguish between CG and GC.

In order to build modular repressors, it is far easier to arrange an additional helix binding to DNA than to work with arms wrapping around DNA. The present invention discloses ways to vary and select correct recognition in these non-helical interactions, but in the preferred embodiment, only helical interactions will be exploited.

Lambda repressor and lambda CRO compete for the same six sites. A single site is 17 base pairs long. A full lambda operator site comprises three copies of this 17-mer; there are two full operators in lambda. Operators of different phages or bacterial repressors are of different lengths in the range of 15-20 base pairs.

Often, a pair of dimeric repressors bind to two identical, or nearly identical, palindromic sequences (which are in proper register) in a cooperative manner, as shown in Fig. 5. This means that binding of repressor to the first site is tighter when a repressor molecule is already bound to the second site than if no repressor is bound to the second site. Similarly, binding of repressor to the second site is tighter when a repressor molecule is already bound to the first site than if no repressor is bound to the first site. Four groups of base pairs are contacted by four identical portions of the proteins. Each of the four regions comprises

five base pairs. The high specificity of repression follows from protein-DNA contact over 20 base pairs. Accidental reproduction of this site should occur only once in 1,099,511,627,776 bases; the human genome is only one thousandth this size. If recognition extends to six regions of five bases, accidental reproduction of this site should occur only once in 1,000,000,000,000,000,000!

In actual repressors, recognition is not absolute; altering one base pair out of ten does not abolish binding, but lowers it by a factor of ten or so. Indeed, each substitution will have a different effect. Thus, if we truly need to bind to only one location in the genome of an organism, we will need to recognize a sequence of a length that would occur only once in a random sequence ten or a hundred times as large as the actual genome. There are sequences in eukaryotic genomes which are highly repetitive. Of course, we cannot repress such sequences, but they are not expressed anyway, so there is no need.

When a repressor binds tightly, transcription of the genes which follow is blocked.

DNA-binding proteins in general, and repressors in specific, must first come in contact with DNA before recognition is possible. Recognition involves exclusively short-range forces DNA is negatively charged. Repressors must be able to approach the DNA without serious hindrance.

On the other hand, excessive positive charge on a repressor leads to high non-specific binding.

Consider a natural repressor which binds DNA both specifically and non-specifically.

Let Kspec = [operator][R]/[R:operator] and let

Kns = [DNA][R]/[R:DNA]

If we build a tetradentate repressor by just doubling a bidentate repressor, we might expect Kspec′ = Kspec*Kspec and Kns′ = Kns*Kns.

This might seem favorable because the ratio of specific to non-specific binding has increased. The kinetics of such a repressor, however, might be quite unsatisfactory, as the repressor will bind too strongly to DNA in general. What is needed is a molecule that binds to eukaryotic DNA as specifically as phage receptors bind to bacterial DNA. Clearly the amount of non-specific binding must be carefully adjusted, and some of the added specific interactions must be used to compensate for reduced non-specific binding.

Lambda repressor comprises a dimer, each dimer contains two domains. The amino-terminal domain binds to DNA; the carboxy-terminal domain dimerizes. There is a 38-amino-acid linker between the N- and C-domains. The amino domains do not dimerize or bind DNA at physiological concentrations if cut from C-domains. The C-domains do dimerize even if cut off from N-domains.

Intact dimers of lambda repressor bind to adjacent operator sites cooperatively. This cooperation arises from interactions between the C-dimers of different repressor dimers.

In Fig. 6, a flow chart describing the steps of the present invention is shown.

In step 2000, a natural dimeric sequence-specific DNA-binding protein is selected. A variety of such repressors are known. Although knowledge of the three-dimensional structure is very helpful, use of modeling and protein-sequence homology will allow one to use some repressors for which no actual three-dimensional structure is available.

In Fig. 7, part a) shows two DNA-binding domains of lambda CRO joined (genetically) by a protein linker. As the recognition helices are now sequentially linked, they can be changed independently. Part b) shows a CRO dimer in which the interface has been mutated so that unlike monomers will associate in preference to like monomers. Part c) shows a

molecule such as lambda repressor in which the dimerization domain (Dx and Dx-tilde) will associate heterologously. The recognition domains in c) will differ only in their recognition components.

In step 2010, an asymmetric bidentate repressor is produced. As illustrated in Fig. 7, either an amino acid linker is introduced to convert the dimeric molecule into a single chain or the dimer contact is modified to make asymmetric dimers much more stable than symmetric dimers.

Although many methods could be used to select the required linker for the first method, the preferred embodiment involves selection of a protein sequence by an extension of the methods of U.S. Patent Application Serial No. 902,970, incorporated herein by reference. The gaps to be closed may be quite large. This leads to uncertainty in design of the linker; generation of a large population of related DNA sequences which code for a population of potential repressors followed by in vivo selection will find a suitable protein sequence so that the pseudo dimer folds correctly and binds sequence-specifically to DNA. A direct protein design method could also be used.

Making asymmetric dimer contacts, the second method can be achieved genetically as long as the sequence of the gene is known. If the domain structure is known, genetic variation can be focused more carefully so that the desired asymmetry can be introduced more easily. If a three-dimensional structure is known, asymmetry can be introduced quite easily.

Asymmetric dimeric association can also be achieved by connecting DNA-binding domains (which do not dimerize by themselves) to protein domains which are known to form heterodimers, e.g., the light and heavy Fv domains or the S-peptide and an inactivated ribonuclease domain.

In step 2020, a dictionary containing at least one protein sequence for the DNA-binding region for each of the

1024 possible DNA five-base sequences is established. Below a method involving in vivo selection from a population created by intentionally-varied in vitro DNA synthesis is given, but other methods could be used.

In step 2030, an asymmetric tridentate repressor is developed as illustrated in Fig. 8. Starting from one of the asymmetric bidentate repressors developed in step 2010, an additional recognition element is engineered so that it will contact the major groove of DNA at a definite place relative to the binding sites of the first two recognition elements. There are three possibilities:

a) connect a new recognition element to an asymmetric dimeric CRO-like molecule,

b) connect a new recognition element to an asymmetric dimeric lambda-repressor-like molecule, and

c) connect a new recognition element to an asymmetric single-chain repressor molecule.

In step 2040, an asymmetric tetradentate repressor is developed as illustrated in Fig. 9. Eight avenues are open:

a) two pseudo-dimer molecules are linked to produce a single-chain tetradentate molecule (1 chain),

b) an additional recognition element can be attached to a single-chain tridentate repressor (1 chain),

c) connect new recognition elements to either side of an asymmetric dimeric CRO-like molecule (2 chains),

d) connect new recognition elements to either side of an asymmetric dimeric lambda-repressor-like molecule (2 chains),

e) the tetramerization interface of the C-domains of lambda-repressor can be made asymmetric (4 chains),

f) asymmetric dimerization domains can be attached to two different single-chain bidentate repressors (2 chains),

g) asymmetric dimerization domains can be attached to two different asymmetric dimeric bidentate repressors (4 chains), and

h) asymmetric dimerization domains can be attached to one asymmetric dimeric bidentate repressor to a single-chain bidentate repressor (3 chains).

The protein linkers required for the conversions in steps a, b, c, and d may be selected in many ways, but in the preferred embodiment a method involving selection of a protein sequence by an extension of the methods of U.S. Patent Application Serial No. 902,970 is given. Linking two molecules into one establishes a definite relationship between the positions on the DNA to which the recognition elements of the first repressor molecule bind and the positions on the DNA to which the new recognition element of the new domain bind. Again generation of a large population of related DNA sequences which code for a large population of potential tetradentate repressors, followed by in vivo selection, will produce a protein sequence which will bind with high sequence specificity to DNA.

Generation of asymmetric dimeric interfaces in steps e, f, g, and h can be achieved genetically. Structural information can be used to focus genetic variation to the amino acids in the interface.

In step 2050, a variety of pentadentate repressors are produced as illustrated in Fig. 10. The methods described in step 2040 can be used by obvious extension.

In step 2060, a variety of hexadentate repressors are created. The methods described in step 2040 can be used by obvious extension.

In step 2080, frameworks are designed and produced which respond to chemical signals. Using one of the repressors developed in steps 2010, 2030, 2040, 2050, or 2060 and the recognition elements developed in step 2020, an essential gene

of E. coli (or some other suitable bacteria) is repressed. The linker or interface region or regions is or are varied when the gene is made by in vitro DNA synthesis. This population is transfected into E. coli or some other suitable bacteria. First a selection is performed which eliminates unrepressed cells. This must be done because the variation in the framework could produce nonfunctional repressors; such selections are well-known in bacterial genetics. Now the slowly growing bacteria are supplied with the chemical which will be the signal. Any cells which grow are likely to have one or more binding sites for the chemical signal. Binding the chemical messenger could cause a conformational change in the repressor which abolishes DNA binding.

This strategy could be reversed. First grow the cells, thereby selecting for repressors which fail to bind DNA in the absence of the selected chemical message. Now introduce the chemical message and select against fast-growing cells by standard genetic means.

The chemical message could be an unusual ion or small molecule. Possible examples are barium ions, DDT, and tetrabromobiphenyl.

Summary of Framework Creation and Dictionary Creation

In steps 2010, 2030, 2040, 2050, and 2060, a series of repressor frameworks were produced with different numbers of DNA-binding elements. In step 2080, frameworks were selected which are sensitive to specific chemical signals. Once the repressor frameworks are determined, the DNA-binding elements can be changed to obtain binding to different DNA sequences.

When repression of a particular gene in a particular cell class is needed, one can calculate the level of specificity needed and some of the steps may be omitted. For example, if repression is sought in a bacterial system, tridentate binding

would be adequate, and steps 2040, 2050, and 2060 could be omitted. In a eukaryotic system, higher specificity is usually needed. If tetradentate binding is sufficiently specific, one can choose to do steps 2010 and 2040. If chemical signals are not needed to control repression, step 2080 may be omitted.

In step 2090, an organism is selected in which we wish to repress one or more genes.

Steps 2100, 2110, 2120, and 2130 are repeated for each target DNA sequence for which a repressor is sought in the selected organism.

In step 2100, a target DNA sequence is selected. The DNA sequence need not have any particular symmetry. The selection criteria would include a) sufficient length to obtain required specificity, b) checks against accidental appearance in locations which are not to be repressed, and c) location in the genome. It has been observed that repressor bound to an operator in the middle of a gene reduces the amount of DNA transcription following the operator site. The reduction is, however, much less than if the repressor had bound in such a way the RNA polymerase could have never bound. Thus, the best place to bind a repressor is over a promoter. If the promoter is shared with a gene which should not be repressed, we might need to choose a DNA sequence downstream. This will be less effective, but by placing two or more repressors together, we can reduce gene expression as much as is needed.

In step 2110, specificity components, selected from the dictionary established in step 2020, are substituted into the two, three, four, five, or six slots on the selected repressor framework.

In step 2120, the complete gene for the repressor is constructed with appropriate controls (promoters, self regulation, etc.). For example, a repressor might shut off

its own synthesis once sufficient repressor is present in the cell.

In step 2130, the complete gene for the repressor is tested for repression. If needed, closely related variants are produced and the best repressor is selected in vivo.

In step 2140, the complete gene is introduced into the target cell. The target may be a bacterial, plant, animal, or fungal cell.

Steps 2090 through 2140 are repeated for each target organism.

A more detailed description of each of the above steps is given below.

Fig. 11 illustrates steps 2000 and 2010 in greater detail.

Many natural DNA-binding proteins have been identified by standard biochemical methods. A natural dimeric protein which binds to a specific DNA sequence is selected in step 3000. Criteria for this choice are:

1. Knowledge of the DNA and protein sequence,
2. Knowledge of the DNA sequences to which it binds,
3. Knowledge of related repressors with different operator sequences,
4. Knowledge of domain structure of protein, and
5. Knowledge of the three-dimensional structure of the protein.

Through genetic engineering, one can now produce enough repressor for sequencing and for crystallography. If a dimeric protein is to be linked into a single chain, a three-dimensional model must be obtained by X-ray crystallography, NMR spectroscopy, or by modeling from the three-dimensional structure of a related repressor, sequence homology, and theoretical methods. The structure need not be of very high accuracy; the portion of the molecule which contains the DNA must be correctly identified; the envelope of the molecule

must be approximately correct; and the disposition of the carboxy and amino terminii must be correct. The more accurate the model, the more readily will the following steps be accomplished. An X-ray structure determined at 3.0 Angstroms would be adequate so long as the course of the protein main chain was correct.

If one intends to use asymmetric association (dimers or tetramers), it is sufficient to know the sequence and be able to identify the DNA-binding components. As will be evident from later parts, the more we know about the natural repressor, the easier things will go. To use asymmetrical association (dimers or tetramers), one must know at least two related repressors with different operator sequences.

In step 3010, the DNA-binding components of the natural repressor are identified. This identification may be by any one or a combination of standard methods, such as:

a) Modeling from X-ray or NMR structure of the repressor,

b) Elimination of X-ray structure of a complex between repressor and DNA,

c) Genetics, and

d) Chemistry (such as methylation masking).

In all repressors studied so far, the DNA-binding component consists of an alpha helix of nine amino acids, plus the amino acid one before the helix. A few other amino acids have small effects on the binding, but these can be ignored until a later stage. All proteins which use an alpha helix to contact the major groove of DNA in the same orientation will share the same recognition dictionary described in step 2020, to a high degree of approximation. Of course, one must remember that the alpha helix can run along the groove in either direction and that the helix can rotate. In repressors cI-lambda, CRO-lambda, and, as indicated by sequence homology, almost all other repressors rotation of the recognition helix within the

major groove of DNA is r stricted by the existence of a preceding alpha helix which sits roughly at right angles to the recognition helix. This helix would collide with the DNA should the recognition helix rotate.

Ways to adjust for slight perturbations which different frameworks will impose on the recognition helices are discussed below.

Should one select a natural DNA-binding protein which uses something other than an alpha helix as a recognition component, one must repeat step 2020 for the new kind of recognition component. Should one select a new DNA-binding protein which uses alpha helices, one might be able to reuse the dictionary determined in step 2020 for some different natural protein, but one might need to repeat step 2020.

Asymmetric repressors are divided into two classes: single-chain and heterodimers.

The first step in producing a single-chain general gene repressor of high specificity is to convert the dimeric DNA-binding domain of a natural repressor to a single chain.

To this end, any DNA-binding protein might be used; positive regulators can easily be made into repressors if the amino acids which bind RNA polymerase are modified to be non-functional, such mutants occur in nature. Converting a two-chain aggregate to a single-chain can be achieved using the methods of U.S. Patent Application Serial No. 902,970 or any equivalent method.

In step 3020, the methods of Application Serial No. 902,970 or some equivalent method are used to design a linker which will convert the natural dimeric DNA-binding protein to a single-chained pseudo-dimer. The methods of Application Serial No. 902,970 will produce a definite sequence for a single-chained pseudo-dimer repressor. To optimize the structure of this pseudo-dimer, an expert first identifies, in step 3030, regions of the design which are least plausible.

In step 3040, the gene for the designed pseudo-dimer is synthesized with intentional infidelity in the uncertain regions. This population of potential repressor genes is introduced into a suitable bacterial host and expressed. In addition, a gene which is normally repressed by the natural dimeric repressor and which is highly deleterious to the host is introduced. Those pseudo-dimeric repressors which fold correctly and bind to the natural palindromic operator will repress the deleterious gene and cells containing those genes will live. Repressors which do not fold will not function, and cells containing those genes will die. If the gene for the putative repressor is under control of a known inducible repressor, then the level of expression of putative repressor can be controlled.

For example, one could put a putative lambda repressor under control of the lac repressor. Lac repression is alleviated by isopropylthiogalactoside (IPTG). A population of E. coli carrying a population of putative lambda repressors under control of lac repressor are infected by a cI mutant of lambda. Those E. coli carrying and suppressing functional repressor genes will survive. The best repressor can be selected by adjusting the concentration of IPTG.

In step 3060, the sequence of the pseudo-dimer is refined by variation around the sequence selected in step 3050. Residues which an expert judges to be suboptimal from

1) X-ray structure,
2) NMR,
3) modeling, and
4) comparison of DNA sequences of different survivors of steps 3050

are varied by intentionally unfaithful in vitro DNA synthesis followed by in vivo selection. Should we find that an amino acid which was varied in step 3040 is the same in all survivors of step 3050, then we can assume that it is optimal,

given its surroundings. By examining our 3-D model, we can see which amino acids are in contact with the non-constant residues determined from multiple isolates. These amino acids should be varied in the next round of selection.

Production of a heterodimeric repressor requires slightly different kinds of information than are needed to produce a single-chain asymmetric repressor. While production of a single-chain repressor required a three-dimensional model of the repressor, so that linkers could be designed, only a single palindromic operator sequence was needed. To produce a heterodimeric repressor, no three-dimensional structure is required (though great use can be made of one). Instead, one requires at least two different palindromic sequences and the protein sequences of the corresponding DNA-recognition elements.

The following example will make this clear. The phages lambda and 434 have similar repressors which bind different DNA sequences as given in the table.

DNA sequence	protein sequence of helix alpha 3	phage
	1 2 3 4 5 6 7 8 9	
ATCAC	gln-ser-ala-ile-asn-lys-ala-ile-his	lambda cI
CAAGA	gln-gln-ser-ile-glu-gln-leu-glu-asn	434R cI

The X-ray structure of lambda repressor and sequence homology indicate that residues 4 and 7 are away from the DNA. The small amino acids (SER and ALA) at residue 3 probably do not contact DNA. Residue 8 may or may not contact the DNA.

Now, in step 3070, construct a plasmid containing the following six genes (shown in Fig. 12):

1) lambda cI gene under control of lac operator,
2) lambda cI gene(a3=434) under control of trp operator,

3) a highly deleterious gene controlled by a hybrid lambda-434 operator,

4) a beneficial (e.g. his+) gene controlled by a lambda-lambda operator,

5) a beneficial (e.g. tyr+) gene controlled by a 434-434 operator, and

6) a gene for drug resistance.

The second gene has part of the DNA-recognition helix alpha 3 replaced by the 434 sequence; residues 4 and 7 are left as the lambda sequence, the rest are from 434. Transfect E. coli or some other suitable bacteria with this plasmid. Let the host bacteria be deficient for genes 4 and 5 (his-, tyr- in the example given).

Under selective pressure from the drug, cells without the plasmid will die. Unless the trp and lac repressors are turned off, cells with the plasmid will die because the deleterious gene is not repressed. If trp and lac repressors are turned off, the two lambda-like repressors will be expressed. Because the ability to dimerize is located in the C-domain, a binomial distribution of repressors will form:

2 [434-lambda]::1 [lambda-lambda]::1 [434-434].

The deleterious gene will be repressed, as will genes 4 and 5. The cells will grow if his and tyr are supplied, but slowly if his and tyr are limiting.

Now, in step 3080, introduce mutations into the C-domains of both lambda-like repressors. The C-domain of lambda repressor contains 104 residues, judging from other proteins which aggregate, 10% to 20% of these residues will be involved in the dimer interface. Each residue on one side of the interface will touch two or three residues on the other side.

The lambda system is highly evolved, so one might assume that the dimer interface is well optimized. Consider a mutation which changes an amino acid in the interface of the C-domain of gene 1. This change will occur in both chains and

so will have double effect. Most mutations in the interface will make the dimer less stable. This mutation will have only a single effect on the hybrid dimer (lambda-434). If the destabilization of lambda-lambda is moderate, mass action will cause the concentration of hybrid dimers to increase. The repression of gene 4 would be reduced and the repression of gene 3 increased.

Now consider the case that a mutation in the C interface of gene 1 touches a mutation in the C interface of gene 2. This is in fact the case we want.

If we assume that there are ten residues in the interface, then each residue in the interface touches 30% of the residues on the other side of the interface. If we put k changes in each C-domain, then in $(0.1)(0.1)(0.3)(k)(k) = 0.003k^2$ of the cases, we will have a mutation in C1 touching a mutation in C2.

Thus, our strategy is to make hydrophobic groups in C1 bigger and change the sign of charged groups in C1 at the same time that we make hydrophobic groups in C2 smaller and change the sign of charged groups. In a second round of mutations, we make C1 smaller and C2 bigger. We could just introduce random changes, and the same pattern would emerge, but more slowly. It may be sufficiently easier so that this will be the best way.

Consider the following recipe for making genes for C-domain.

Amino Acid Given	Bigger & change sign: Codon Used	Bigger & change sign: Amino Acids Obtained	Smaller & change sign: Codon Used	Smaller & change sign: Amino Acids Obtained
ALA	G (C+xT) T	ALA + xVAL	G (C+xG) T	ALA + xGLY
CYS	T (G+xA) C	CYS + xTYR	(T+xG) G C	CYS + xGLY
ASP	(G+xA) A C	ASP + xASN	(G+xA) A C	ASP + xASN
GLU	(G+xA) A A	GLU + xLYS	(G+xA) A A	GLU + xLYS
PHE	T (T+xA) C	PHE + xTYR	(T+xC) T C	PHE + xLEU
GLY	G (G+xC) T	GLY + xALA	G G T	GLY
HIS	(C+xT) A C	HIS + xTYR	(C+xA) A C	HIS + xTYR
ILE	(A+xT) T C	ILE + xPHE	(A+xG) T C	ILE + xVAL
LYS	(A+xG) A A	LYS + xGLU	(A+xG) T C	LYS + xGLU
MET	(A+xC) T G	MET + xLEU	(A+xG) T G	MET + xVAL
ASN	(A+xC) A C	ASN + xHIS	A (A+xG) C	ASN + xSER
PRO	C (C+xT) G	PRO + xLEU	(C+dA) C G	PRO + xTHR
GLN	(C+xG) A G	GLN + xGLU	C A (G+xC)	GLN + xHIS
ARG	C (G+xA) T	ARG + xHIS	C (G+xA) T	ARG + xHIS
SER	A (G+xC) C	SER + xTHR	(A+xG) G C	SER + xGLY
THR	A (C+xA) C	THR + xASN	A (C+xG) C	THR + xSER
VAL	(G+xT) T T	VAL + xPHE	G (T+xC) T	VAL + xALA
TRP	T G G	TRP	T (G+xT) G	TRP + xLEU
TYR	T A C	TYR	(T+xC) A C	TYR + xHIS

If x is selected so that each amino acid is about 97% correct, then about 5% of each C-domain will be the protein sequence written. 95% of the C-domains will have between 1 and 6 changes. The construction given greatly increases the chance that heterodimers will become much more stable than homodimers.

Because the two homodimers control different genes, we can find out which homodimer is less repressed.

Sequencing the survivors will tell us where changes were made. If multiple changes are introduced, we will need to do them one at a time by site-directed mutagenesis to see which

are truly in interface. From these experiments, a genetic map of what touches what in C-domain will emerge.

Once a natural dimeric repressor is converted to an asymmetric bidentate repressor (i.e., a molecule in which the DNA-binding components are related by an approximate diad, but there are other parts which are not diad related), the amino acids in either of the recognition components may be changed independently. Because the recognition components may now be different, the repressor can be made to bind to non-palindromic sequences.

It has already been demonstrated that a definite relationship exists between the five bases of DNA to be recognized and the sequences of the part of each protein which fits into the major groove. That is, we can build a table of length 1024 in which there will be a protein sequence (of some defined length, much less than the whole protein) for each possible pentadeoxynucleotide. For example, the DNA binding component of lambda repressor is a nine-amino-acid alpha helix. The following table relates a few DNA sequences to the sequences of DNA-binding helices.

DNA Sequence	Protein Sequence	Phage
ATCAC	gln-ser-ala-ile-asn-lys-ala-ile-his	lambda cI
TTTAA	glu-ser-gln-ile-ser-arg-trp-lys-gly	F22 cI
CAAGA	gln-gln-ser-ile-glu-gln-leu-glu-asn	434R cI

There will be a plurality of possible protein sequences corresponding to each possible pentadeoxynucleotide; it is sufficient, however, that for each pentanucleotide there exists one protein sequence which will bind more strongly to this DNA sequence than to any other DNA sequence. A preferred method of establishing this table will be given below, but other methods could be used. (One may discover that interactions with bases either side of the main four are of some importance, in which case the table will be of length 4096 or 16384, but the methods given will allow one to fill this table also.) Parts of this table of 1024 relationships are already known from studies of natural repressors.

The dictionary of recognition elements will be developed as shown in Fig. 13. In step 4010, 1024 versions of the plasmid shown in Fig. 12 will be created which have hybrid operators. One-half of each operator will have the natural DNA sequence. The other side of the operator will be systematically varied so that every possible pentanucleotide sequence will appear in the positions of major recognition.

In step 4020, the gene for the asymmetric bidentate repressor will be synthesized <u>in vitro</u> so that one of the recognition helices will be the natural sequence. The other recognition helix will be highly varied to produce a wide population of recognition helices. The population of genes will be introduced into a bacterial host and expressed. The

hybrid operators are positioned to repress a highly deleterious gene. Thus one can select in vivo (step 4030) working repressors which bind to each of the hybrid operators. The dictionary is compiled simply by sequencing the repressor gene of the winner in each of the 1024 in vivo selections.

One method of varying the recognition helix is as follows. Beginning one residue before the recognition helix (studies show that this residue plays a part in recognition), synthesize the gene as follows:

Position in Helix	Observed		Codon			Range substituted
-1	GYSTKDNEGAR		-T	-T	any	PHRGTNSKADEG
		or	T	any	C	FSYC (to get Y)
1	GNDRVYKEPIS		-T	any	any	PHRSTNSKADEGLIMV
		or	T	any	C	FSYC (to get Y)
2	SVAGELP		-T	any	any	PHRGTNSKADEGLIMV
3	GASTGRN		-T	-T	any	PHRGTNSKADEG
4	VILA		-T	T+C	any	LPITMVA
5	GNSEGRYT		-T	-T	any	PHRGTNSKADEG
		or	T	any	C	FSYC (to get Y)
6	AKGSRVWRE		-T	any	any	PHRGTNSKADEGLIMV
		or	T	any	C	FSYC (to get Y)
		or	T	G	G	W
7	LAWIVHGYE		-T	any	any	PHRSTNSKADEGLIMV
		or	T	any	C	FSYC (to get Y)
		or	T	G	G	W
8	FIGKELVCSH		-T	any	any	PHRGTNSKADEGLIMV
		or	T	any	C	FSYC (to get Y and C)
		or	T	G	G	W
9	NHREAKGPL		-T	any	any	PHRGTNSKADEGLIMV

Completely random DNA for ten codons yields $4^{30} = 10^{20}$, a very large number. This will code for about 10^{13} different protein sequences, including stop. The scheme given above reduces the number of possible protein sequences by 20-fold and avoids stop codons. More restrictive recipes could be devised.

Using the methods of Application Serial No. 902,970, one can connect a small domain of protein containing at least one alpha helix of nine or more amino acids and lying on the surface to one of the asymmetric bidentate repressors developed in steps 3020 through 3060 or steps 3070 through 3090. (This section needs a flow chart.) The new domain must be connected so that the selected alpha helix will lie in the major groove of the DNA one or more base pairs removed from the helices of the bidentate unit. An alpha helix lying in the major groove of DNA retains two important degrees of freedom: a) translation along its own axis, and b) rotation about its own axis. These motions will determine how the amino acids will interact with the base pairs. If our model places the new helix in the same rotational and translational relationship to the DNA as is observed by the pseudo-dimer, then we might be able to use the same dictionary of recognition elements as determined in step 2020. If we do not preserve this relationship, a separate dictionary will be

needed for the added helix. Although this is additional work, it will function and is a valid method.

Assume that the alpha helix of the new domain will contact the DNA in the same way as the helices of the pseudo-dimer. The sequence of lambda operator and a model of the potential tridentate repressor will show which base pairs of the operator will contact the new helix. From the dictionary determined in step 4030 one selects the appropriate protein sequence for the recognition helix.

From a model of a tridentate repressor an expert identifies amino acids which might not be optimal and the gene for this repressor is synthesized with intentional infidelity for those identified amino acids. This population of potential tridentate repressors is subjected to in vivo selection.

Tetradentate repressors can be made from single chains or by careful asymmetric aggregation.

To obtain sufficiently high specificity, the methods of U.S. Patent Application Serial No. 902,970 or equivalent methods will be used to generate a single polypeptide chain which will resemble two pseudo-dimers insofar that the DNA-recognizing components will be correctly positioned to contact the major groove of B-DNA. This novel protein will be tetradentate and will have a specificity of one in 10,000,000,000 base pairs.

Using the dictionary described above in steps 2020 and 4030, we can take a given DNA sequence and select sequences for each of the four recognition components which will cause the novel repressor to bind strongly to the desired DNA sequence.

Just as in vivo selection was used to refine the linkers which created a single-chained pseudo-dimer from the natural dimeric DNA-binding protein, the linkers which join two pseudo-dimers into a tetradentate repressor can be refined by in vivo selection. First an operator of known sequence is positioned before a deleterious gene. A model of the tetradentate repressor will show which parts of the DNA will contact each of the DNA-recognizing elements of the tetradentate repressor. The correct protein sequences for each recognition element is found in the dictionary of recognition elements (step 2020 and 4030). The parts of the linkers which are least certain are identified by an expert and those parts of the gene for the tetradentate repressor are made intentionally unfaithful. This produces a population of potential tetradentate repressors which vary in the linkage which positions the four DNA-recognizing elements. This population is subjected to in vivo selection.

Repressors with five or six recognition helices can be made by repeating the steps used to generate asymmetric bidentate, tridentate, and tetradentate repressor frameworks.

In the preferred embodiment, the CRO (lambda) repressor is used, but any other repressor for which an X-ray structure or other acceptable three-dimensional structure is available can be used. The single-chained pseudo-dimer contains one copy of the DNA-binding domain, a suitable linker, and a second copy of the DNA-binding domain. The recognition helices of the CRO (lambda) repressor are replaced by those of the lambda repressor.

In the case of CRO (lambda) repressor, the C-terminal arm (residues 61-67) is assumed to wrap around the DNA. The linker is built by an extension of U.S. Patent Application Serial No. 902,970 as follows. From the C-terminal of chain 1 to the amino-terminal of chain 2 is a wide groove where the two chains come in contact. An expert user viewing this structure on computer graphics decided that an alpha helix would lie in this groove and make many favorable contacts. In addition, a helical linker would make internal hydrogen bonds and so be self-stabilizing. Thus, alpha helices were extracted from a protein in the Brookhaven Protein Data Bank. In this case, the A, B, and E helices of the alpha chain and the H helix of the beta chain were taken from human deoxy haemoglobin. These helices were selected because they were not too hydrophobic; the dimer contacts of CRO repressor are not very hydrophobic, but contain several arginines, and eglutamic acids. All of these helices have one or more basic residues

(lysine or arginine) near the amino-terminus. Because the C-terminus of chain 1 is near the DNA and the amino-terminus of the other chain is far from the DNA, the alpha helix must run away from the DNA (i.e., N-terminus is near DNA and C-terminus is far). In other cases (e.g., lambda repressor), the overall chain direction is reversed and the amino-terminal arm wraps around DNA and the C-terminal lies far from the DNA. This would cause one to select other alpha helices with basic residues near the C-terminal. If a natural dimeric repressor had its C-terminal near enough to the N-terminal of the other chain, one would dispense with helix to span most of the gap and instead look directly for linkers.

Each of these four helices from haemoglobin was laid in the intermolecular groove in turn. Using the methods of U.S. Patent No. 4,704,692, short linkers from the C-terminal of CRO to the alpha helix, and from the alpha helix to the N-terminal of the other chain were found. Given the distance to be spanned and the number of amino acids needed, there is some uncertainty that the exact sequence selected will fold correctly and bind the correct DNA sequence with high specificity. Thus, our preferred method is to use an <u>in vivo</u> selection to find the linker sequence which will allow the molecule to fold.

This selection proceeds as follows. The DNA sequence of CRO repressor is prepared for residues 1 to 63 except that

the specificity helix is that of lambda repressor. The DNA of the linker is made with intentional infidelity so that only a small percentage of the final chain has exactly the specified sequence. As the sequence of the linker contains 20 to 30 amino acids, a completely random DNA sequence would generate 21^N sequences (many of which would terminate prematurely) which is far too great to sample by <u>in vivo</u> selection. For a sequence of 30 amino acids, all possible point mutations gives 570 additional sequences; all possible double mutations gives 1.5×10^5 sequences; all possible triple mutations gives 2.7×10^7 sequences; and quadruple mutations adds 3.5×10^{10} mutations. We wish, therefore, to vary the protein sequence in the "neighborhood" of the given sequence.

Examination of the genetic code allows us to generalize protein sequences as follows:

Amino acid specified	Codon used	Amino acids obtained
A	G (C+G) any	A G
C	(T+A) G (T+C)	C S
D	-T -T any	P T A H Q N K D E R S G
E	-T -T any	P T A H Q N K D E R S G
F	(T+A) (T+A) (T+C)	F Y I N
G	G (C+G) any	A G
H	-T -T any	P T A H Q N K D E R S G
I	any T any	F L I M V
K	-T -T any	P T A H Q N K D E R S G
L	any T any	F L I M V
M	any T any	F L I M V
N	-T -T any	P T A H Q N K D E R S G
P	-T -T any	P T A H Q N K D E R S G
Q	-T -T any	P T A H Q N K D E R S G
R	-T -T any	P T A H Q N K D E R S G
S	-T -T any	P T A H Q N K D E R S G
T	-T -T any	P T A H Q N K D E R S G
V	any T any	F L I M V
W	(T+A) (T+G) G	W L M R
Y	T (T+A) (T+C)	Y F

Columns 1 & 3 use the single-letter amino acid code:

A:ala	C:cys	D:asp	E:glu	F:phe	G:gly	H:his	I:ile
K:lys	L:leu	M:met	N:asn	P:pro	Q:gln	R:arg	S:ser
T:thr	V:val	W:trp	Y:tyr				

Column 2 uses the notation -T for (A,C,G)

A more restrictive scheme would be

Amino acid specified	Codon used	Amino acids obtained
A	G (C+G) any	A G
C	(T+A) G (T+C)	C S
D	-T A any	H Q N K D E
E	-T A any	H Q N K D E
F	T (T+Q) (T+C)	F Y
G	G (C+G) any	A G
H	-T A any	H Q N K D E
I	any T any	F L I M V
K	-T A any	H Q N K D E
L	any T any	F L I M V
M	any T any	F L I M V
N	-T A any	H Q N K D F
P	C -T any	P H Q R
Q	(C+A) -T any	P T H Q N K R S
R	-T (A+G) any	H Q N K D E R S G
S	A -T any	T N K R S
T	A -T any	T N K R S
V	any T any	F L I M V
W	(T+A) (T+G) G	W L M R
Y	T (T+A) (T+C)	Y F

A yet more restrictive scheme would be

Amino acid specified	Codon used	Amino acids obtained
A	G (C+G) any	A G
C	(T+A) G (T+C)	C S
D	G A any	D E
E	G A any	D E
F	T (T+G) (T+C)	F Y
G	G (C+G) any	A G
H	-G A (T+C)	H Y N
I	-T T any	L I M V
K	A -T (A+G)	T K R
L	any T any	F L I M V
M	any T G	L M V
N	(C+A) A any	H Q N K
P	C -T any	P H Q R
Q	(C+A) A any	H Q N K
R	(C+A) (A+G) any	H Q N K R S
S	A -T any	T N K R S
T	A -T any	T N K R S
V	any T any	F L I M V

W	(T+A) (T+G) G	W L M R
Y	T (T+A) (T+C)	Y F

Using the restrictive scheme, the number of sequences grows something like 5 raised to the number of amino acids.

Our target is 10^8 sequences; we should remember that

$$10^8 \approx 2^{25} \approx 5^{11} \approx 20^6$$

where the equations are only approximate.

If we have only two choices at each location, we can use random sequences for 26 amino acids. If there are five choices, we can run 11 amino acids, and if full substitution is needed, only 6 amino acids can be varied.

Once these two domains are incorporated into a single polypeptide chain, we will be free to alter the amino acids which contact the DNA. Because the two regions of DNA binding are quite separate and because we have covalently linked them, we can change them independently. In this way, we will escape the requirement of a palindromic DNA sequence.

We can generate different sequence specificity by:

(a) examining published data on protein sequence vs. bound DNA sequence;

(b) protein engineering from crystal structures; and

(c) mutagenesis of those residues which contact DNA.

WE CLAIM:

1. A gene repressor which comprises two or more sequence-specific DNA-binding domains covalently joined by a polypeptide.

2. A recombinant DNA molecule encoding the gene repressor of claim 1.

3. A method of treating viral infections in an individual by administering to an individual infected with a virus a virus growth inhibiting amount of the molecules of claim 1.

4. A method of preventing viral infections in an individual by administering to said individual a viral growth inhibiting amount of the molecules of claim 1.

5. A gene repressor which comprises two or more sequence-specific DNA-binding molecules wherein each DNA-binding molecule possesses at least one region capable of specifically binding to a different DNA-binding molecule.

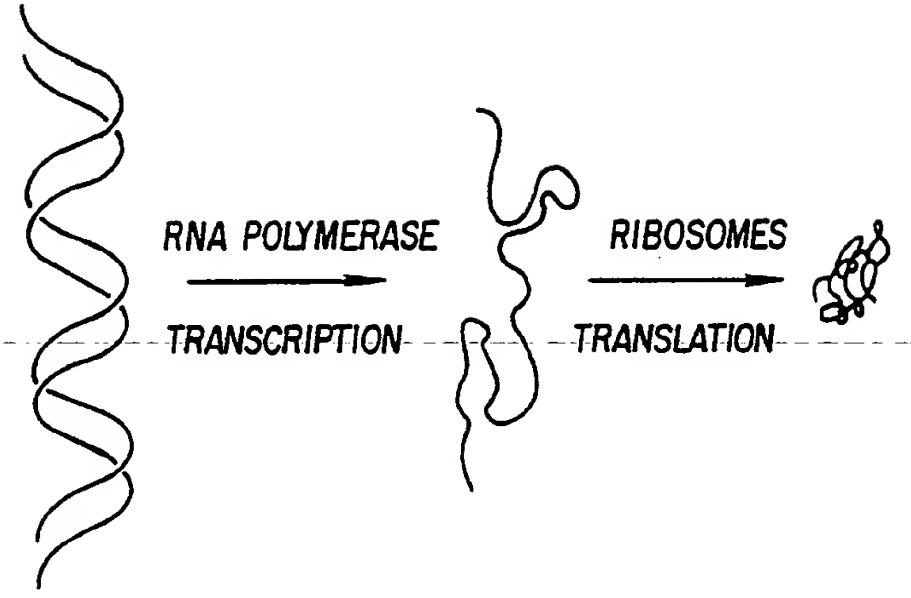
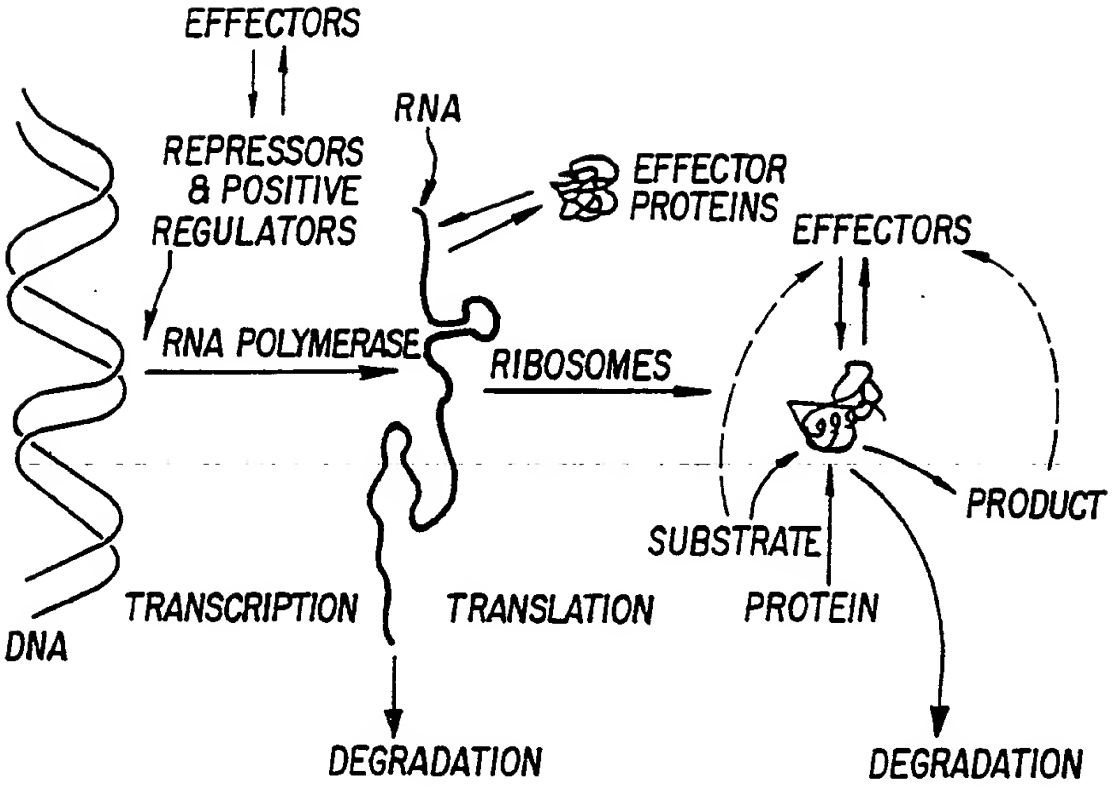
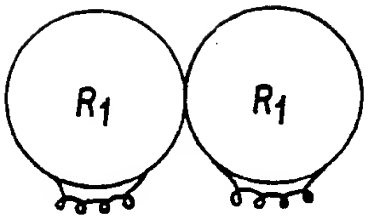

RNA POLYMERASE
TRANSCRIPTION
RIBOSOMES
TRANSLATION
DNA
RNA
Protein
TRANSCRIPTION & TRANSLATION

FIG. 1

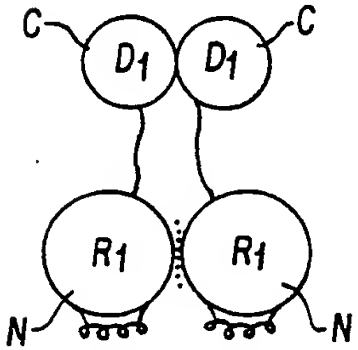


CELLULAR REGULATION

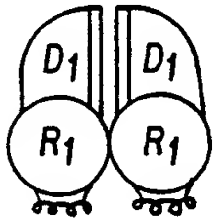
FIG. 2



CRO- λ



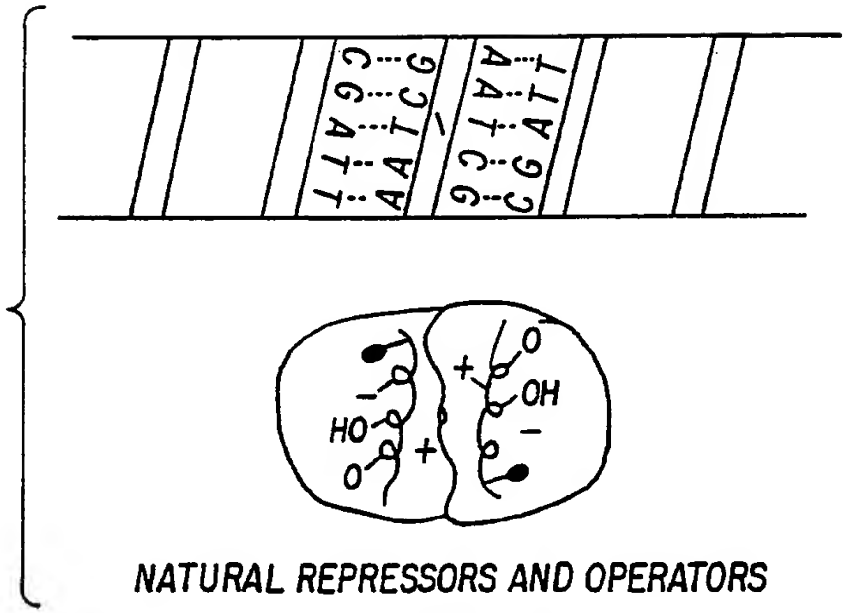
λ REPRESSOR



CAP

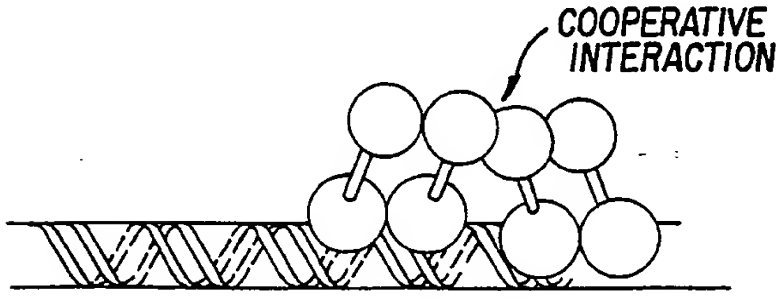
NATURAL DNA-BINDING PROTEINS

FIG. 3



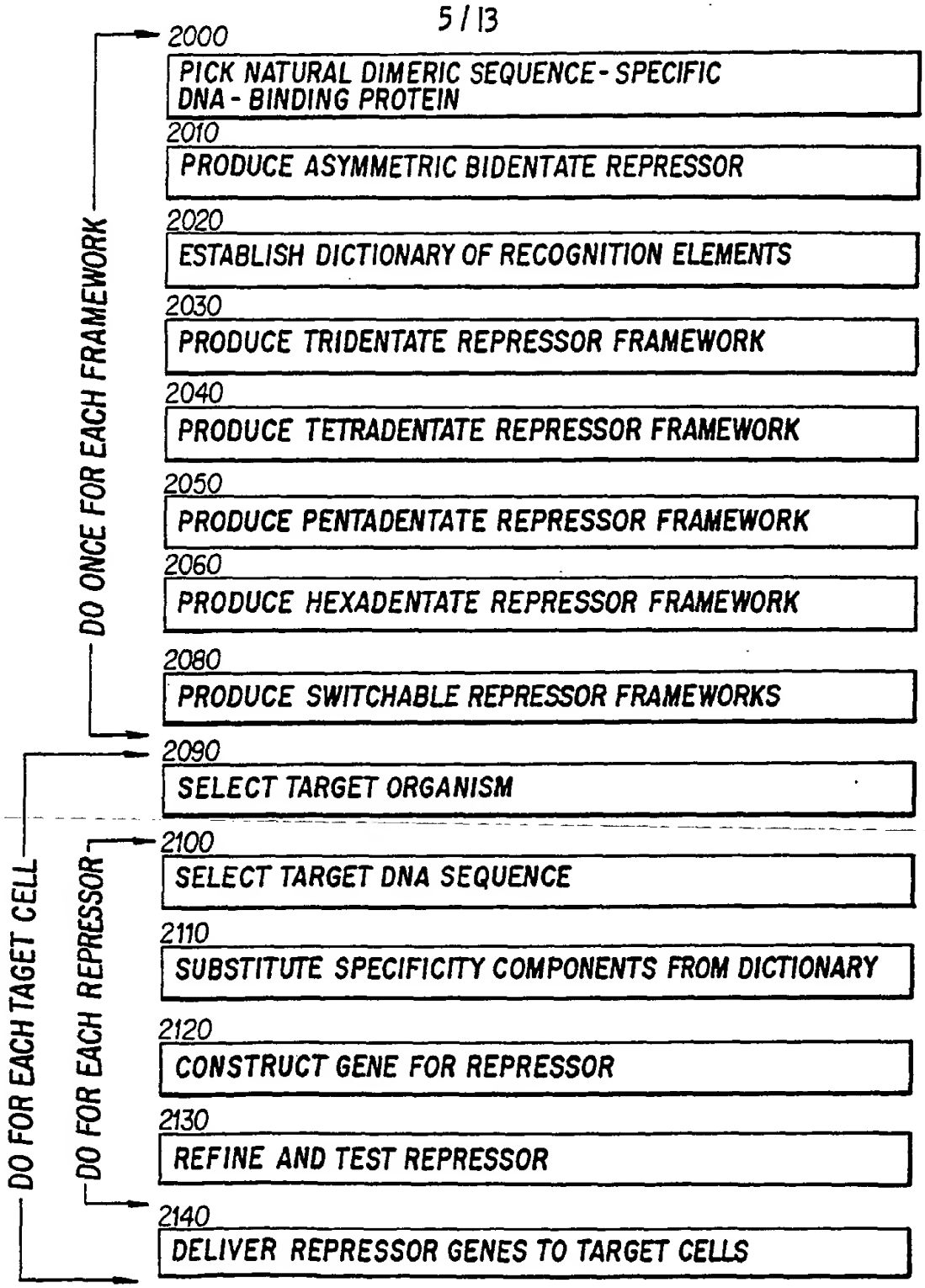
NATURAL REPRESSORS AND OPERATORS

FIG. 4



COOPERATIVE BINDING OF LAMBDA REPRESSOR

FIG. 5



DEVELOPMENT OF GENERAL GENE REPRESSOR

FIG. 6

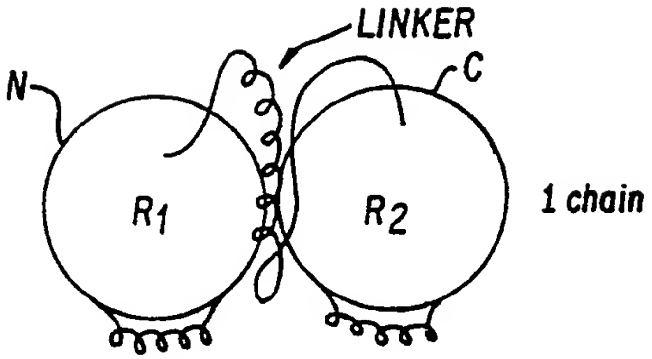


FIG. 7(a)

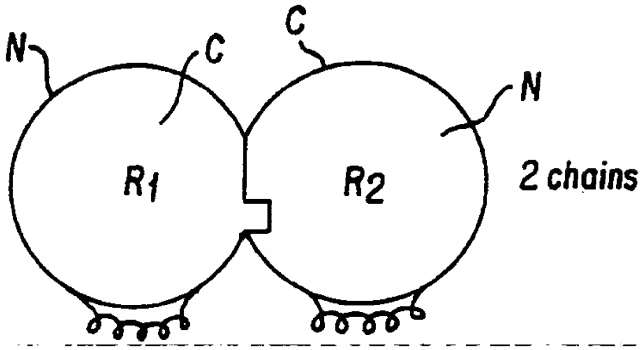


FIG. 7(b)

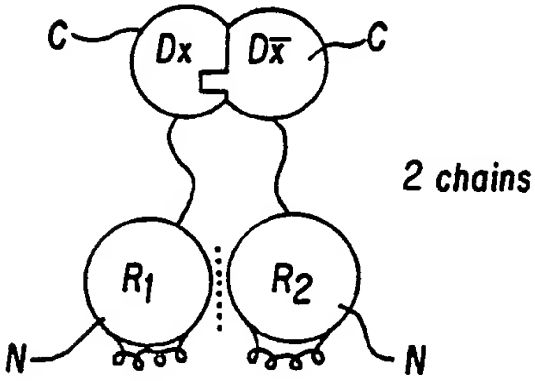


FIG. 7(c)

SCHEMATIC ASYMMETRIC BIDENTATE REPRESSORS

FIG. 8(a)

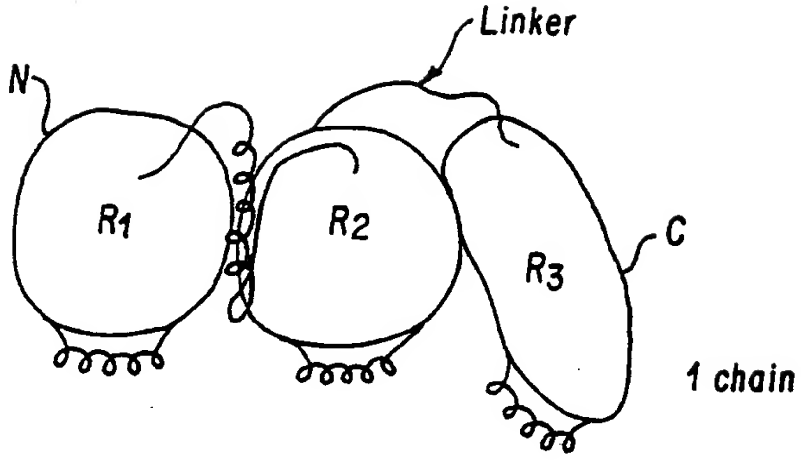


FIG. 8(b)

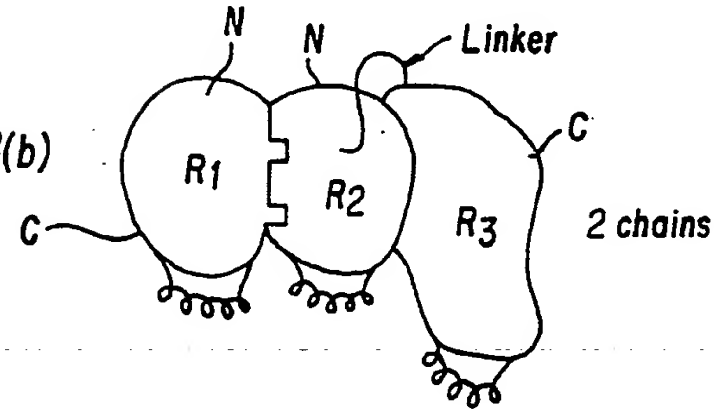
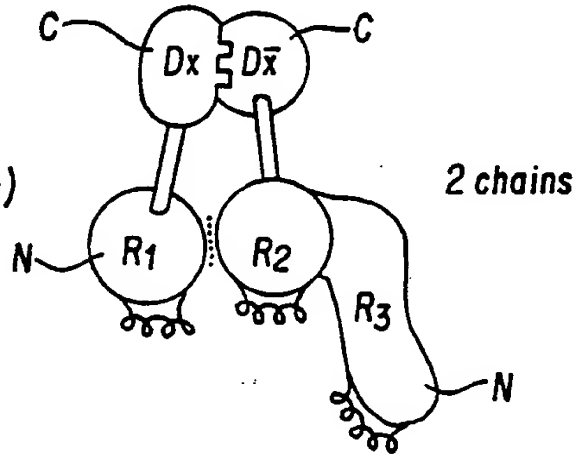
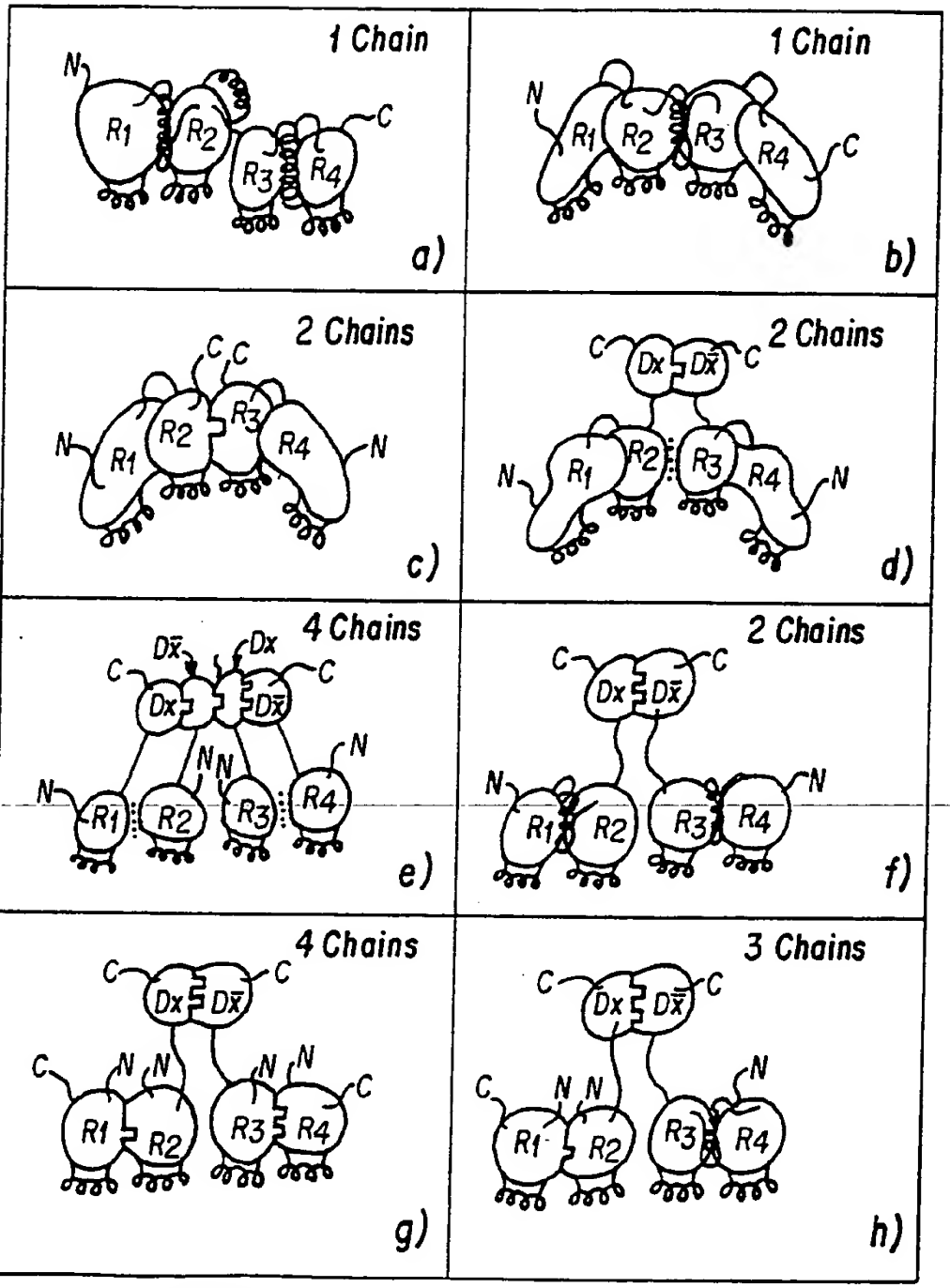


FIG. 8(c)

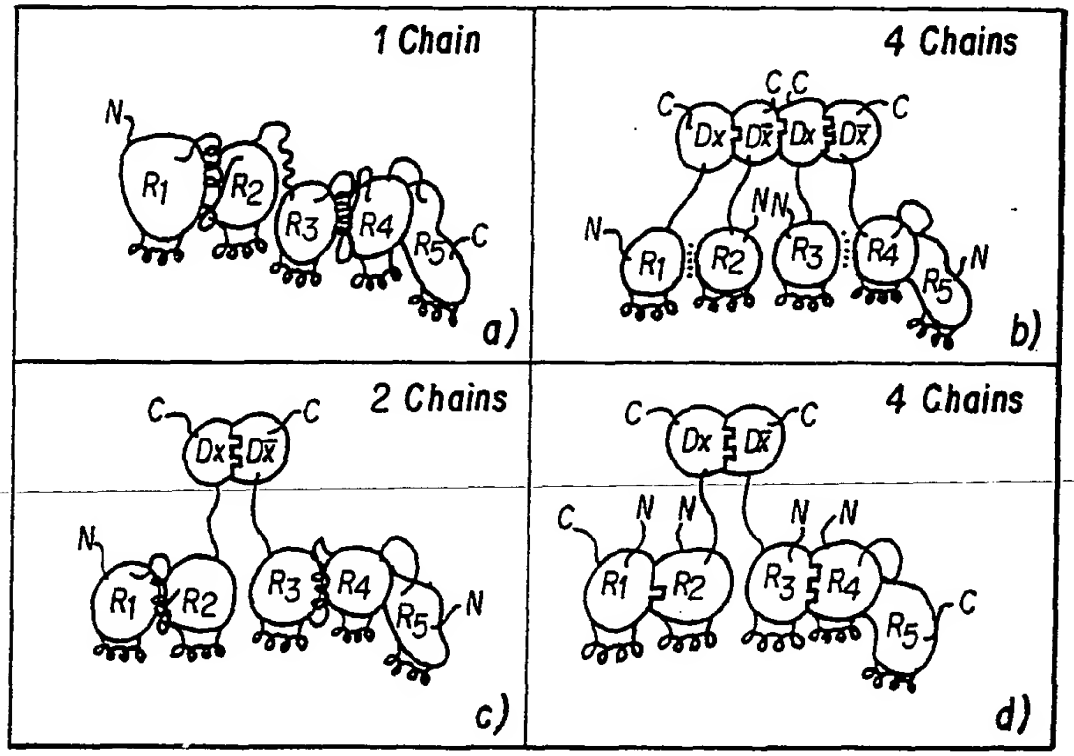


SCHEMATIC TRIDENTATE REPRESSOR FRAMEWORKS



SCHEMATIC TETRADENTATE REPRESSOR FRAMEWORKS

FIG. 9



SCHEMATIC PENTADENTATE REPRESSORS

FIG. 10

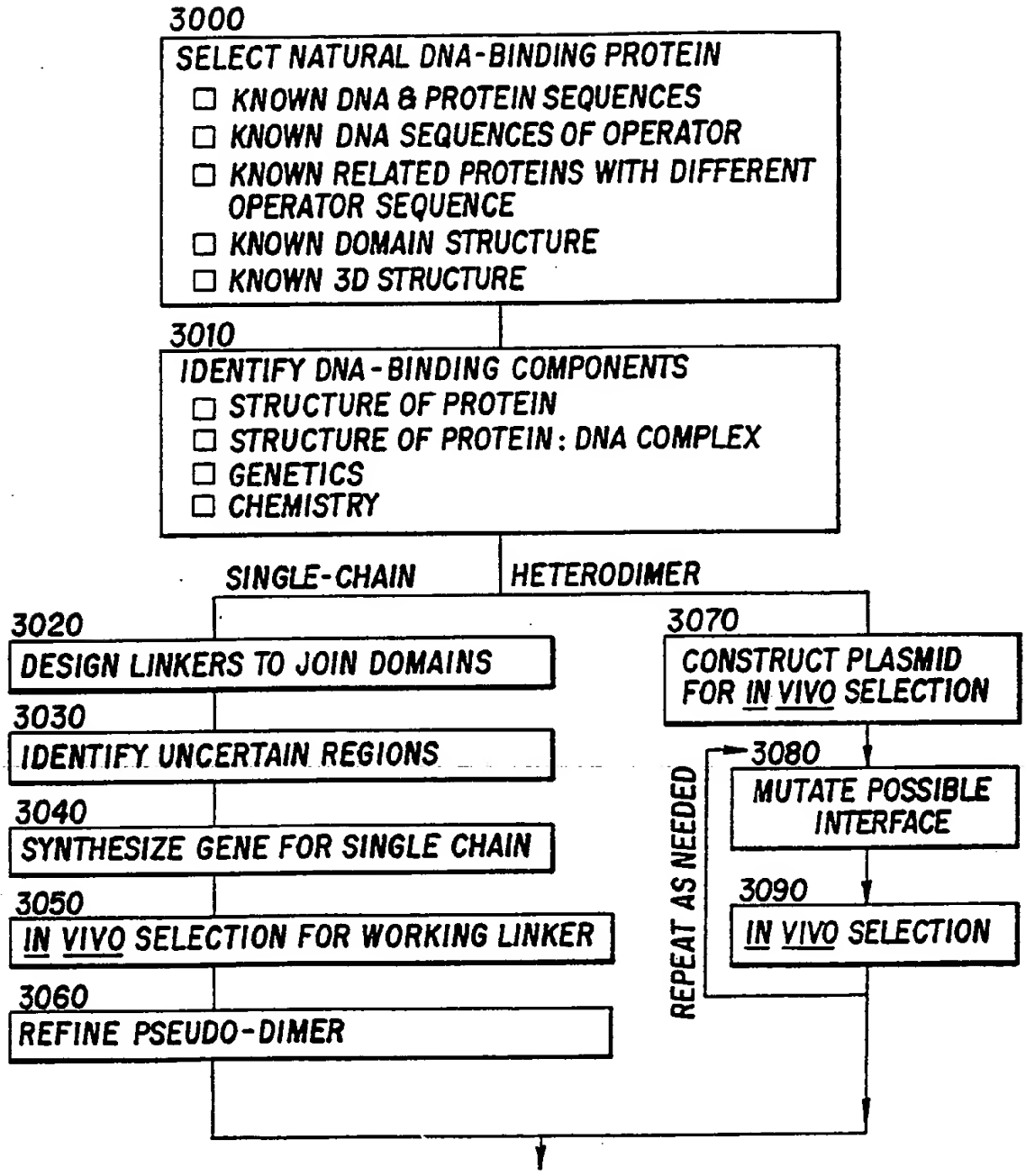


FIG. 11 NATURAL & CREATED BIDENTATE DNA BINDERS

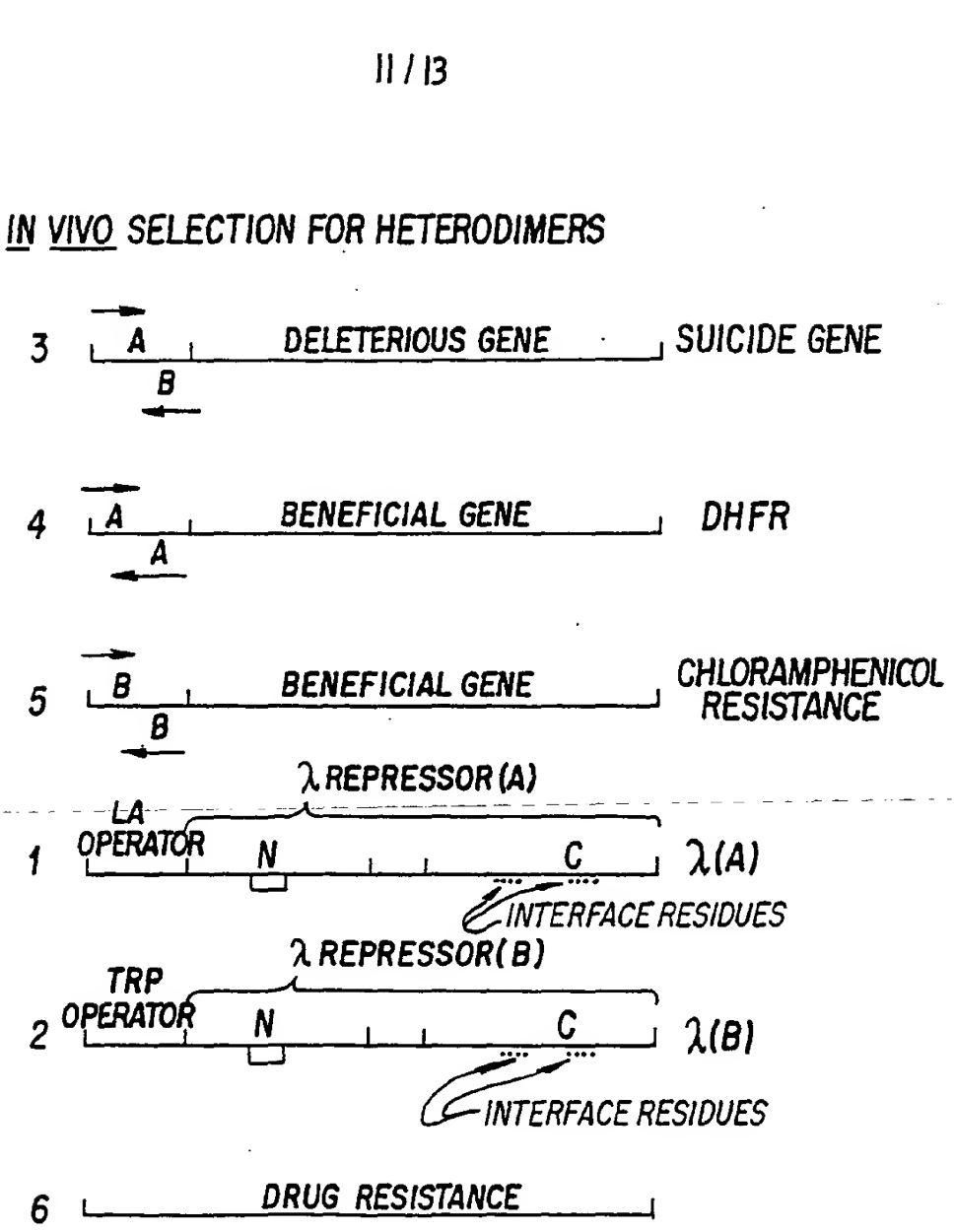

IN VIVO SELECTION FOR HETERODIMERS
3
A
B
DELETERIOUS GENE
SUICIDE GENE
4
A
A
BENEFICIAL GENE
DHFR
5
B
B
BENEFICIAL GENE
CHLORAMPHENICOL RESISTANCE
λ REPRESSOR (A)
LA
OPERATOR
1
N
C
λ(A)
INTERFACE RESIDUES
λ REPRESSOR (B)
TRP
OPERATOR
2
N
C
λ(B)
INTERFACE RESIDUES
6
DRUG RESISTANCE

FIG. 12

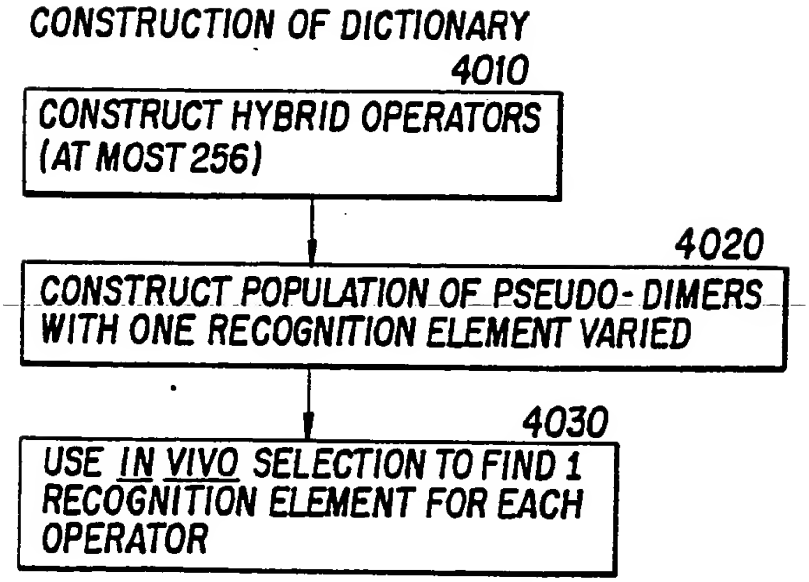

CONSTRUCTION OF DICTIONARY
4010
CONSTRUCT HYBRID OPERATORS (AT MOST 256)
4020
CONSTRUCT POPULATION OF PSEUDO-DIMERS WITH ONE RECOGNITION ELEMENT VARIED
4030
USE IN VIVO SELECTION TO FIND 1 RECOGNITION ELEMENT FOR EACH OPERATOR

FIG. 13

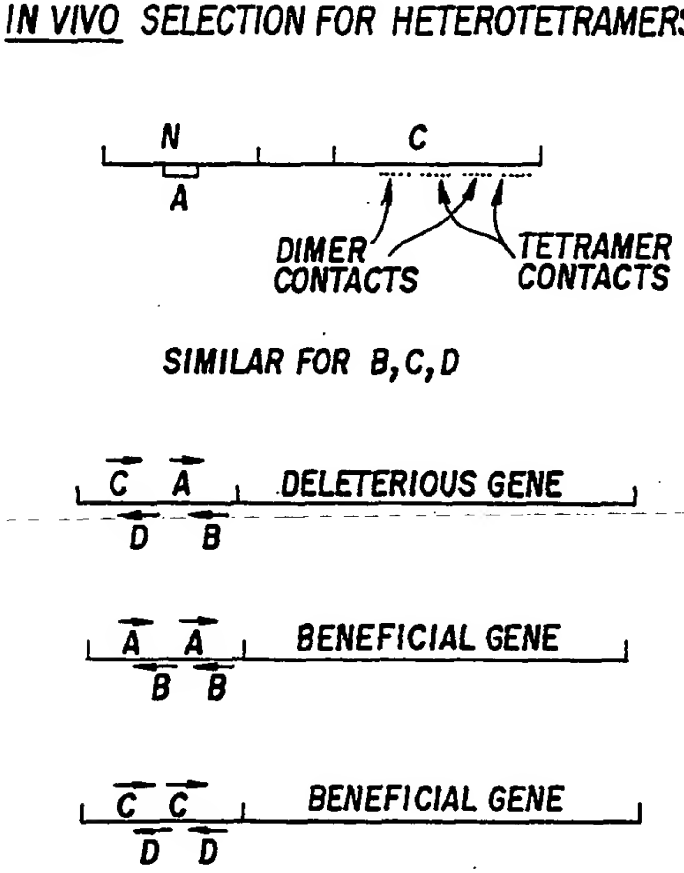

IN VIVO SELECTION FOR HETEROTETRAMERS
N
C
A
DIMER CONTACTS
TETRAMER CONTACTS
SIMILAR FOR B, C, D
C A
DELETERIOUS GENE
D B
A A
BENEFICIAL GENE
B B
C C
BENEFICIAL GENE
D D

FIG. 14

INTERNATIONAL SEARCH REPORT

International Application No PCT/US88/00718

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) ³

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC(4): C07K 15/00; C07H 15/12, 17/00; A61K 37/02
US CL : 530/350; 536/27; 536/28; 536/29; See Attachment

II. FIELDS SEARCHED

Minimum Documentation Searched ⁴

Classification System ⁵	Classification Symbols
U.S.	530/350; 536/27; 536/28; 536/29; 514/2; 514/21; 935/6; 935/11; 935/40 .

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched ⁶

CHEMICAL ABSTRACTS FILE "CA" AND BIOSIS DATA BASE SEARCHED WITH TERM "GENE(W) REPRESS? AND DNA(IA) BIND?"

III. DOCUMENTS CONSIDERED TO BE RELEVANT ¹⁴

Category ⁷	Citation of Document, ¹⁶ with indication, where appropriate, of the relevant passages ¹⁷	Relevant to Claim No. ¹⁸
X	D. Freifelder, "Molecular Biology", published 1983, by Jones and Bartlett, Inc., A comprehensive Introduction to Prokaryotes and Eukaryotes, (Boston, MA), see pages 561-570.	1,2 and 5
X	Proc. Nat. Acad. Sci. USA, Volume 77, issued 1980, (Washington, D.C. U.S.A.), Schmeissner et al., Promoter for the establishment of repressor synthesis in bacteriophage. See pages 3191-3195.	1,2,5
X	Proc. Nat. Acad. Sci. USA, Volume 73, issued 1976, (Washington, D.C. U.S.A), Backman et al., Construction of plasmids carrying the cI gene of bacteriophage , See page 4174-4178.	1,2,5

⁷ Special categories of cited documents: ¹⁵

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search ²	Date of Mailing of this International Search Report ³
13 June 1988	28 JUN 1988
International Searching Authority ¹	Signature of Authorized Officer ¹⁹
ISA/US	L.Eric Crane

Form PCT/ISA/210 (second sheet) (October 1981)

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category *	Citation of Document, [14] with indication, where appropriate, of the relevant passages [15]	Relevant to Claim No [18]
X	The EMBO Journal, Volume 4, issued 1985, (Oxford, England), Guilfoyle et al., Two functions encoded by adenovirus early region 1A are responsible for the activation and repression of the DNA-binding protein gene, see pages 707-713.	1,2,5
X	Cell, Volume 45, issued 1986, (U.S.A.), Zinn et al, Detection on Factors That Interact with the Human Interferon Regulatory Region In Vivo by DNAase I Footprinting, see pages 611-618.	1,2,5
X	Nucleic Acids Research, Volume 15, issued 1987, (Oxford, England), Grossman et al., Purification and DNA binding properties of the blaI gene product, repressor for the lactamase gene, blaP, of Bacillus licheniformis, see pages 6049-6062.	1,2,5
X	Proc. Nat. Acad. Sci., U.S.A., Volume 82, issued 1985, (U.S.A), Isackson et al, Dominant negative mutations in the Tn10 tet repressor: Evidence for use of the conserved helix-turn-helix motif in DNA binding, see pages 6226 - 6230.	1,2,5
X	Molecular and Cellular Biology, Volume 5, issued 1985 (U.S.A), Brady et al, trans Activation of the Simian Virus 40 Late Transcription Unit by T-Antigen, see pages 1391-1399.	1,2,5
A	J. Mol. Biol., Volume 162, issued 1982, (London, England), Sanger et al, Nucleotide Sequence of Bacteriophage DNA, see pages 729-773.	1,2,5

PCT/US88/00718

Attachment to Form PCT/ISA/210, Part I.

US CL: 514/2; 514/21;
935/6; 935/11; 935/40